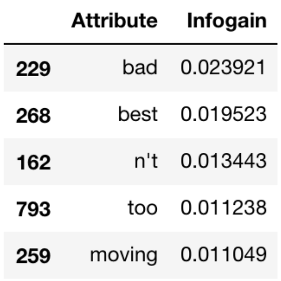
Question 5: Programming Exercise

a. List the 5 tokens that occur most frequently in the training set.

```
.       1426.0
,        816.0
the      780.0
and      737.0
a        722.0
```

b. Using just the training documents, calculate the information gain of every attribute. List the 5 attributes with the highest information gain.



	Attribute	Infogain
229	bad	0.023921
268	best	0.019523
162	n't	0.013443
793	too	0.011238
259	moving	0.011049

c. Report your results on the test set using a confusion matrix. Also, list the percentage accuracy obtained.
Parameters: Hidden Nodes = 2, Learning Rate = 0.01, Epochs = 1000

```
Confusion Matrix
True Positive = 231     False Negative = 42
False Positive = 113    True Negative = 114
--------------------------------------------------------------------------
Neural Network Accuracy (In Percentage) = 69.0
```

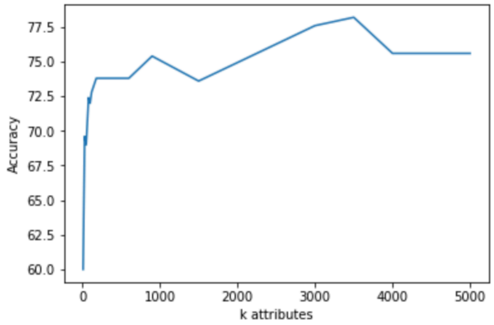
d. What percentage accuracy is achieved if you use Zero-R, instead of a neural net?

```
Zero-R Classifier Accuracy (In Percentage) = 54.6
```

e. We used only the top 50 attributes, but we could have used the top k attributes, for larger or smaller k. Why is it reasonable to think that increasing the number of attributes might increase accuracy? Why is it reasonable to think that decreasing the number of attributes might increase accuracy? Answer both questions.

As we increase the number of attributes for our neural net, while training the data we're making the model more adaptive and it starts learning more details, this helps in increasing the accuracy to a point, but if we increase furthermore, we see that our model has learned even very small details which may lead to overfitting because it decreases the generalization of the model. So, the increase in attributes helps in increasing the accuracy of the model to a certain point but after that, it may lead to overfitting resulting.

f. Perform experiments to see how accuracy changes as you vary the number k of attributes for your neural net. Choose at least 4 values of k (in addition to k = 50) and graph the results. The horizontal axis should correspond to the number of attributes, and the vertical to the test accuracy. Did the results surprise you? Did you have any difficulties running the experiments? Give the graph AND the answers to these two questions.



So from the graph we can see that the accuracy is increasing as we increase the number of attributes (till k < 1500) but at certain points it decreases (like k = 1500) and then again increases (at k = 3000) and then decrease at (k=3500), it does not keep on increasing continuously but shows dips at certain points and then again start increasing. No, we didn't face many difficulties in running the experiment, but it required more time for performing a neural net.